



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification⁶:

G06F 9/38, 9/34

A1

(11) International Publication Number:

WO 99/50740

(43) International Publication Date:

7 October 1999 (07.10.99)

(21) International Application Number: PCT/US99/04718

(22) International Filing Date: 3 March 1999 (03.03.99)

(30) Priority Data:

09/053,002

31 March 1998 (31.03.98)

US

(71) Applicant (for all designated States except US): INTEL CORPORATION [US/US]; 2200 Mission College Boulevard, Santa Clara, CA 95052 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): ROUSSEL, Patrice [FR/US]; 2034 N.E. 37th Avenue, Portland, OR 97212 (US). THAKKAR, Ticky [GB/US]; 150 S.W. Moonridge Place, Portland, OR 97225 (US). ABDALLAH, Mohammad, A. [-/US]; 104 Austin Drive, Folsom, CA 95630 (US). PENTKOVSKI, Vladimir [RU/US]; 221 Luna Circle, Folsom, CA 95630 (US). COKE, James [US/US]; 4453 Barnett Ranch Road, Shingle Springs, CA 95628 (US).

(74) Agents: TAYLOR, Edwin, H. et al.; Blakely, Sokoloff, Taylor & Zafman LLP, 7th floor, 12400 Wilshire Boulevard, Los Angeles, CA 90025 (US).

(81) Designated States: AL, AM, AT, AT (Utility model), AU (Petty patent), AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, CZ (Utility model), DE, DE (Utility model), DK, DK (Utility model), EE, EE (Utility model), ES, FI, FI (Utility model), GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SK (Utility model), SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

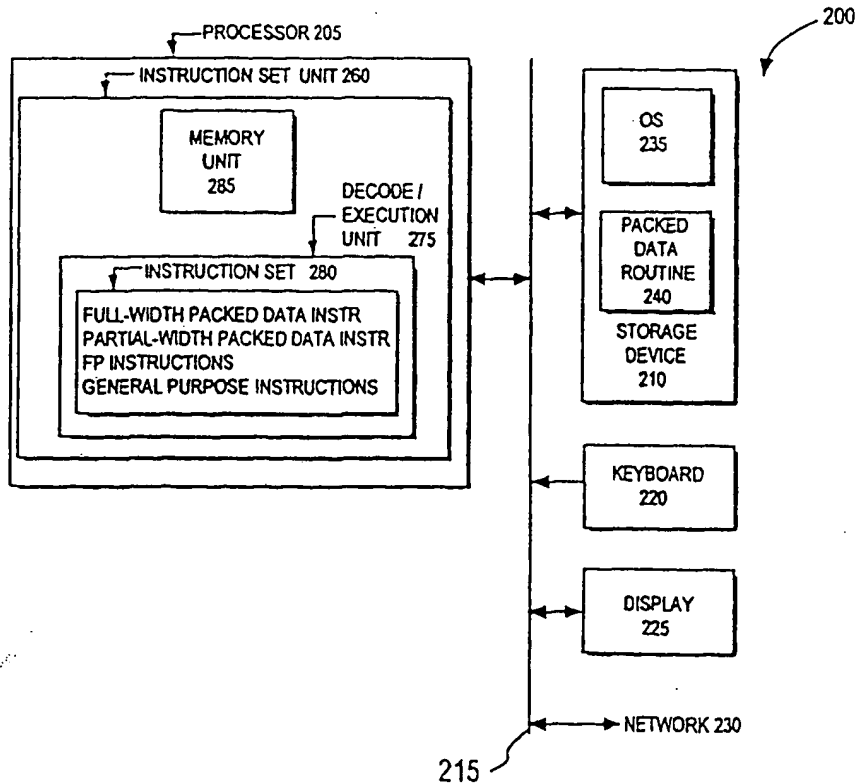
Published

With international search report.

(54) Title: EXECUTING PARTIAL-WIDTH PACKED DATA INSTRUCTIONS

(57) Abstract

A method and apparatus provides for executing scalar packed data instructions. A processor includes a plurality of registers (291, 292), renaming unit coupled to the plurality of registers, and decoder coupled to the register renaming unit. The register renaming unit provides an architectural register file to store packed data operands which include a plurality of data elements. The decoder is configured to decode a first and second set of instructions (e.g., set of full-width packed data instructions and a set of partial width packed data instructions) that specify one or more registers in the architectural register file. Instructions in the first set of instructions specify operations to be performed on all of the data elements stored in one or more specified registers. The instructions in the second set of instructions specify operations to be performed on only a subset of data element stored in the one or more specified registers.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakhstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

EXECUTING PARTIAL-WIDTH PACKED DATA INSTRUCTIONS

FIELD OF THE INVENTION

The invention relates generally to the field of computer systems. More particularly, the invention relates to a method and apparatus for efficiently executing partial-width packed data instructions, such as scalar packed data instructions, by a processor that makes use of SIMD technology, for example.

BACKGROUND OF THE INVENTION

Multimedia applications such as 2D/3D graphics, image processing, video compression/decompression, voice recognition algorithms and audio manipulation, often require the same operation to be performed on a large number of data items (referred to as "data parallelism"). Each type of multimedia application typically implements one or more algorithms requiring a number of floating point or integer operations, such as ADD or MULTIPLY (hereafter MUL). By providing macro instructions whose execution causes a processor to perform the same operation on multiple data items in parallel, Single Instruction Multiple Data (SIMD) technology, such as that employed by the Pentium® processor architecture and the MMx™ instruction set, has enabled a significant improvement in multimedia application performance (Pentium® and MMx™ are registered trademarks or trademarks of Intel Corporation of Santa Clara, CA).

SIMD technology is especially suited to systems that provide packed data formats. A packed data format is one in which the bits in a register are logically divided into a number of fixed-sized data elements, each of which represents a separate value. For example, a 64-bit register may be broken into four 16-bit elements, each of which represents a separate 16-bit value. Packed data instructions may then separately manipulate each element in these packed data types in parallel.

Referring to Figure 1, an exemplary packed data instruction is illustrated. In this example, a packed ADD instruction (e.g., a SIMD ADD) adds corresponding data elements of a first packed data operand, X, and a second packed data operand, Y, to produce a packed data result, Z, i.e., $X_0+Y_0=Z_0$, $X_1+Y_1=Z_1$, $X_2+Y_2=Z_2$, and $X_3+Y_3=Z_3$. Packing many data elements within one register or memory location and employing parallel hardware execution allows SIMD architectures to perform multiple operations at a time, resulting in significant performance improvement. For instance, in this example, four individual results may be obtained in the time previously required to obtain a single result.

While the advantages achieved by SIMD architectures are evident, there remain situations in which it is desirable to return individual results for only a subset of the packed data elements.

SUMMARY OF THE INVENTION

A method and apparatus are described for executing partial-width packed data instructions. According to one aspect of the invention, a processor includes a plurality of registers, a register renaming unit coupled to the plurality of registers, and a decoder coupled to the register renaming unit. The register renaming unit provides an architectural register file to store packed data operands each of which include a plurality of data elements. The decoder is configured to decode a first and second set of instructions that each specify one or more registers in the architectural register file. Each of the instructions in the first set of instructions specify operations to be performed on all of the data elements stored in the one or more specified registers. In contrast, each of the instructions in the second set of instructions specify operations to be performed on only a subset of the data element stored in the one or more specified registers.

Other features and advantages of the invention will be apparent from the accompanying drawings and from the detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention is described by way of example and not by way of limitation with reference to the figures of the accompanying drawings in which like reference numerals refer to similar elements and in which:

Figure 1 illustrates a packed ADD instruction adding together corresponding data elements from a first packed data operand and a second packed data operand.

Figure 2A is a simplified block diagram illustrating an exemplary computer system according to one embodiment of the invention.

Figure 2B is a simplified block diagram illustrating exemplary sets of logical registers according to one embodiment of the invention.

Figure 2C is a simplified block diagram illustrating exemplary sets of logical registers according to another embodiment of the invention.

Figure 3 is a flow diagram illustrating instruction execution according to one embodiment of the invention.

Figure 4 conceptually illustrates the result of executing a partial-width packed data instruction according to various embodiments of the invention.

Figure 5A conceptually illustrates circuitry for executing full-width packed data instructions and partial-width packed data instructions according to one embodiment of the invention.

Figure 5B conceptually illustrates circuitry for executing full-width packed data and partial-width packed data instructions according to another embodiment of the invention.

Figure 5C conceptually illustrates circuitry for executing full-width packed data and partial-width packed data instructions according to yet another embodiment of the invention.

Figure 6 illustrates an ADD execution unit and a MUL execution unit capable of operating as four separate ADD execution units and four separate

MUL execution units, respectively, according to an exemplary processor implementation of SIMD.

Figures 7A-7B conceptually illustrate a full-width packed data operation and a partial-width packed data operation being performed in a "staggered" manner, respectively.

Figure 8A conceptually illustrates circuitry within a processor that accesses full width operands from logical registers while performing operations on half of the width of the operands at a time.

Figure 8B is a timing chart that further illustrates the circuitry of Figure 8A.

Figure 9 conceptually illustrates one embodiment of an out-of-order pipeline to perform operations on operands in a "staggered" manner by converting a macro instruction into a plurality of micro instructions that each processes a portion of the full width of the operands.

Figure 10 is a timing chart that further illustrates the embodiment described in Figure 9.

Figure 11 is a block diagram illustrating decoding logic that may be employed to accomplish the decoding processing according to one embodiment of the invention.

DETAILED DESCRIPTION

A method and apparatus are described for performing partial-width packed data instructions. Herein the term "full-width packed data instruction" is meant to refer to a packed data instruction (e.g., a SIMD instruction) that operates upon all of the data elements of one or more packed data operands. In contrast, the term "partial-width packed data instruction" is meant to broadly refer to a packed data instruction that is designed to operate upon only a subset of the data elements of one or more packed data operands and return a packed data result (to a packed data register file, for example). For instance, a scalar SIMD instruction may require only a result of an operation between the

least significant pair of packed data operands. In this example, the remaining data elements of the packed data result are disregarded as they are of no consequence to the scalar SIMD instruction (e.g., the remaining data elements are don't cares). According to the various embodiments of the invention, execution units may be configured in such a way to efficiently accommodate both full-width packed data instructions (e.g., SIMD instructions) and a set of partial-width packed data instructions (e.g., scalar SIMD instructions).

In the following detailed description, for purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the invention. It will be apparent, however, to one of ordinary skill in the art that these specific details need not be used to practice the invention. In other instances, well-known devices, structures, interfaces, and processes have not been shown or are shown in block diagram form.

Justification of Partial-Width Packed Data Instructions

Considering the amount of software that has been written for scalar architectures (e.g., single instruction single data (SISD) architectures) employing scalar operations on single precision floating point data, double precision floating point data, and integer data, it is desirable to provide developers with the option of porting their software to architectures that support packed data instructions, such as SIMD architectures, without having to rewrite their software and/or learn new instructions. By providing partial-width packed data instructions, a simple translation can transform old scalar code into scalar packed data code. For example, it would be very easy for a compiler to produce scalar SIMD instructions from scalar code. Then, as developers recognize portions of their software that can be optimized using SIMD instructions, they may gradually take advantage of the packed data instructions. Of course, computer systems employing SIMD technology are likely to also remain backwards compatible by supporting SISD instructions as well. However, the many recent architectural improvements and other factors

discussed herein make it advantageous for developers to transition to and exploit SIMD technology, even if only scalar SIMD instructions are employed at first.

Another justification for providing partial-width packed data instructions is the many benefits which may be achieved by operating on only a subset of a full-width operand, including reduced power consumption, increased speed, a clean exception model, and increased storage. As illustrated below, based on an indication provided with the partial-width packed data instruction, power savings may be achieved by selectively shutting down those of the hardware units that are unnecessary for performing the current operation.

Another situation in which it is undesirable to force a packed data instruction to return individual results for each pair of data elements includes arithmetic operations in an environment providing partial-width hardware. Due to cost and/or die limitations, it is common not to provide full support for certain arithmetic operations, such as divide. By its nature, the divide operation is very long, even when full-width hardware (e.g., a one-to-one correspondence between execution units and data elements) is implemented. Therefore, in an environment that supports only full-width packed data operations while providing partial-width hardware, the latency becomes even longer. As will be illustrated further below, a partial-width packed data operation, such as a partial-width packed data divide operation, may selectively allow certain portions of its operands to bypass the divide hardware. In this manner, no performance penalty is incurred by operating upon only a subset of the data elements in the packed data operands.

Additionally, exceptions raised in connection with extraneous data elements may cause confusion to the developer and/or incompatibility between SISD and SIMD machines. Therefore, it is advantageous to report exceptions for only those data elements upon which the instruction is meant to operate. Partial-width packed data instruction support allows a predictable

exception model to be achieved by limiting the triggering of exceptional conditions to those raised in connection with the data elements being operated upon, or in which exceptions produced by extraneous data elements would be likely to cause confusion or incompatibility between SISD and SIMD machines.

Finally, in embodiments where portions of destination packed data operand is not corrupted as a result of performing a partial-width packed data operation, partial-width packed data instructions effectively provide extra register space for storing data. For instance, if the lower portion of the packed data operand is being operated upon, data may be stored in the upper portion and vice versa.

An Exemplary Computer System

Figure 2A is a simplified block diagram illustrating an exemplary computer system according to one embodiment of the invention. In the embodiment depicted, computer system 200 includes a processor 205, a storage device 210, and a bus 215. The processor 205 is coupled to the storage device 210 by the bus 215. In addition, a number of user input/output devices, such as a keyboard 220 and a display 225 are also coupled to bus 215. The computer system 200 may also be coupled to a network 230 via bus 215. The processor 205 represents a central processing unit of any type of architecture, such as a CISC, RISC, VLIW, or hybrid architecture. In addition, the processor 205 may be implemented on one or more chips. The storage device 210 represents one or more mechanisms for storing data. For example, the storage device 210 may include read only memory (ROM), random access memory (RAM), magnetic disk storage mediums, optical storage mediums, flash memory devices, and/or other machine-readable mediums. The bus 215 represents one or more buses (e.g., AGP, PCI, ISA, X-Bus, EISA, VESA, etc.) and bridges (also termed as bus controllers). While this embodiment is described in relation to a single processor computer system, it is appreciated that the invention may be implemented in a multi-processor computer system. In addition while the

present embodiment is described in relation to a 32-bit and a 64-bit computer system, the invention is not limited to such computer systems.

Figure 2A additionally illustrates that the processor 205 includes an instruction set unit 260. Of course, processor 205 contains additional circuitry; however, such additional circuitry is not necessary to understanding the invention. At any rate, the instruction set unit 260 includes the hardware and/or firmware to decode and execute one or more instruction sets. In the embodiment depicted, the instruction set unit 260 includes a decode/execution unit 275. The decode unit decodes instructions received by processor 205 into one or more micro instructions. The execution unit performs appropriate operations in response to the micro instructions received from the decode unit. The decode unit may be implemented using a number of different mechanisms (e.g., a look-up table, a hardware implementation, a PLA, etc.).

In the present example, the decode/execution unit 275 is shown containing an instruction set 280 that includes both full-width packed data instructions and partial-width packed data instructions. These packed data instructions, when executed, may cause the processor 205 to perform full-/partial-width packed floating point operations and/or full-/partial-width packed integer operations. In addition to the packed data instructions, the instruction set 280 may include other instructions found in existing micro processors. By way of example, in one embodiment the processor 205 supports an instruction set which is compatible with Intel 32-bit architecture (IA-32) and/or Intel 64-bit architecture (IA-64).

A memory unit 285 is also included in the instruction set unit 260. The memory unit 285 may include one or more sets of architectural registers (also referred to as logical registers) utilized by the processor 205 for storing information including floating point data and packed floating point data. Additionally, other logical registers may be included for storing integer data, packed integer data, and various control data, such as a top of stack indication and the like. The terms architectural register and logical register are used

herein to refer to the concept of the manner in which instructions specify a storage area that contains a single operand. Thus, a logical register may be implemented in hardware using any number of well known techniques, including a dedicated physical register, one or more dynamically allocated physical registers using a register renaming mechanism (described in further detail below), etc.. In any event, a logical register represents the smallest unit of storage addressable by a packed data instruction.

In the embodiment depicted, the storage device 210 has stored therein an operating system 235 and a packed data routine 240 for execution by the computer system 200. The packed data routine 240 is a sequence of instructions that may include one or more packed data instructions, such as scalar SIMD instructions or SIMD instructions. As discussed further below, there are situations, including speed, power consumption and exception handling, where it is desirable to perform an operation on (or return individual results for) only a subset of data elements in a packed data operand or a pair of packed data operands. Therefore, it is advantageous for processor 205 to be able to differentiate between full-width packed data instructions and partial-width packed data instructions and to execute them accordingly.

Figure 2B is a simplified block diagram illustrating exemplary sets of logical registers according to one embodiment of the invention. In this example, the memory unit 285 includes a plurality of scalar floating point registers 291 (a scalar register file) and a plurality of packed floating point registers 292 (a packed data register file). The scalar floating point registers 291 (e.g., registers R0-R7) may be implemented as a stack referenced register file when floating point instructions are executed so as to be compatible with existing software written for the Intel Architecture. In alternative embodiments, however, the registers 291 may be treated as a flat register file. In the embodiment depicted, each of the packed floating point registers (e.g., XMM0-XMM7) are implemented as a single 128-bit logical register. It is appreciated, however, wider or narrower registers may be employed to

conform to an implementation that uses more or less data elements or larger or smaller data elements. Additionally, more or less packed floating point registers 292 may be provided. Similar to the scalar floating point registers 291, the packed floating point registers 292 may be implemented as either a stack referenced register file or a flat register file when packed floating point instructions are executed.

Figure 2C is a simplified block diagram illustrating exemplary sets of logical registers according to another embodiment of the invention. In this example, the memory unit 285, again, includes a plurality of scalar floating point registers 291 (a scalar register file) and a plurality of packed floating point registers 292 (a packed data register file). However, in the embodiment depicted, each of the packed floating point registers (e.g., XMM0-XMM7) are implemented as a corresponding pairs of high 293 and low registers 294. As will be discussed further below, it is advantageous for purposes of instruction decoding to organize the logical register address space for the packed floating point registers 292 such that the high and low register pairs differ by a single bit. For example, the high and low portions of XMM0-XMM7 may be differentiated by the MSB. Preferably, each of the packed floating point registers 291 are wide enough to accommodate four 32-bit single precision floating point data elements. As above, however, wider or narrower registers may be employed to conform to an implementation that uses more or less data elements or larger or smaller data elements. Additionally, while the logical packed floating point registers 292 in this example each comprise corresponding pairs of 64-bit registers, in alternative embodiments each packed floating point register may comprise any number of registers.

Instruction Execution Overview

Having described an exemplary computer system in which one embodiment of the invention may be implemented, instruction execution will now be described.

Figure 3 is a flow diagram illustrating instruction execution according to one embodiment of the invention. At step 310, an instruction is received by the processor 205. At step 320, based on the type of instruction, partial-width packed data instruction (e.g., scalar SIMD instruction) or full-width packed data instruction (e.g., SIMD instruction), processing continues with step 330 or step 340. Typically, in the decode unit the type of instruction is determined based on information contained within the instruction. For example, information may be included in a prefix or suffix that is appended to an opcode or provided via an immediate value to indicate whether the corresponding operation is to be performed on all or a subset of the data elements of the packed data operand(s). In this manner, the same opcodes may be used for both full-width packed data operations and partial-width packed data operations. Alternatively, one set of opcodes may be used for partial-width packed data operations and a different set of opcodes may be used for full-width packed data operations.

In any event, if the instruction is a conventional full-width packed data instruction, then at step 330, a packed data result is determined by performing the operation specified by the instruction on each of the data elements in the operand(s). However, if the instruction is a partial-width packed data instruction, then at step 340, a first portion of the result is determined by performing the operation specified by the instruction on a subset of the data elements and the remainder of the result is set to one or more predetermined values. In one embodiment, the predetermined value is the value of the corresponding data element in one of the operands. That is, data elements may be "passed through" from data elements of one of the operands to corresponding data elements in the packed data result. In another embodiment, the data elements in the remaining portion of the result are all cleared (zeroed). Exemplary logic for performing the passing through of data elements from one of the operands to the result and exemplary logic for clearing data elements in the result are described below.

Figure 4 conceptually illustrates the result of executing a partial-width packed data instruction according to various embodiments of the invention. In this example, an operation is performed on data elements of two logical source registers 410 and 420 by an execution unit 440. The execution unit 440 includes circuitry and logic for performing the operation specified by the instruction. In addition, the execution unit 440 may include selection circuitry that allows the execution unit 440 to operate in a partial-width packed data mode or a full-width packed data mode. For instance, the execution unit 440 may include pass through circuitry to pass data elements from one of the logical source registers 410, 420 to the logical destination register 430, or clearing circuitry to clear one or more data elements of the logical destination register 430, etc. Various other techniques may also be employed to affect the result of the operation, including forcing one of the inputs to the operation to a predetermined value, such as a value that would cause the operation to perform its identity function or a value that may pass through arithmetic operations without signaling an exception (e.g., a quiet not-a-number (QNaN)).

In the example illustrated, only the result (Z_0) of the operation on the first pair of data elements (X_0 and Y_0) is stored in the logical destination register 430. Assuming the execution unit 440 includes pass through logic, the remaining data elements of the logical destination register 430 are set to values from corresponding data elements of logical source register 410 (i.e., X_3 , X_2 , and X_1). While the logical destination register 430 is shown as a separate logical register, it is important to note that it may concurrently serve as one of the logical source registers 410, 420. Therefore, it should be appreciated that setting data elements of the logical destination register 430 to values from one of the logical source registers 410, 420 in this context may include doing nothing at all. For example, in the case that logical source register 410 is both a logical source and destination register, various embodiments may take advantage of this and simply not touch one or more of the data elements which are to be passed through.

Alternatively, the execution unit 440 may include clearing logic. Thus, rather than passing through values from one of the logical source registers to the logical destination register 430, those of the data elements in the result that are unnecessary are cleared. Again, in this example, only the result (Z_0) of the operation on the first pair of data elements (X_0 and Y_0) is stored in the logical destination register 430. The remaining data elements of the logical destination register 430 are "cleared" (e.g., set to zero, or any other predetermined value for that matter).

Full-Width Hardware

Figures 5A - 5C conceptually illustrate execution units 540, 560 and 580, respectively, which may execute both full-width packed data and partial-width packed data instructions. The selection logic included in the execution units of Figures 5A and 5C represent exemplary pass through logic, while the selection logic of Figure 5B is representative of clearing logic that may be employed. In the embodiments depicted, the execution units 540, 560, and 580 each include appropriate logic, circuitry and/or firmware for concurrently performing an operation 570, 571, and 572 on the full-width of the operands (X and Y).

Referring now to Figure 5A, the execution unit 540 includes selection logic (e.g., multiplexers (MUXes) 555-557) for selecting between a value produced by the operation 570 and a value from a corresponding data element of one of the operands. The MUXes 555-557 may be controlled, for example, by a signal that indicates whether the operation currently being executed is a full-width packed data operation or a partial-width packed data operation. In alternative embodiments, additional flexibility may be achieved by including an additional MUX for data element 0 and/or independently controlling each MUX. Various means of providing MUX control are possible. According to one embodiment, such control may originate or be derived from the instruction itself or may be provided via immediate values. For example, a 4-bit immediate value associated with the instruction may be used to allow the

MUXes 555-557 to be controlled directly by software. Those MUXes corresponding to a one in the immediate value may be directed to select the result of the operation while those corresponding to a zero may be caused to select the pass through data. Of course, more or less resolution may be achieved in various implementations by employing more or less bits to represent the immediate value.

Turning now to Figure 5B, the execution unit 540 includes selection logic (e.g., MUXes 565-567) for selecting between a value produced by an operation 571 and a predetermined value (e.g., zero). As above, the MUXes 565-567 may be under common control or independently controlled.

The pass through logic of Figure 5C (e.g., MUXes 575-576) selects between a data element of one of the operands and an identity function value 590. The identity function value 590 is generally chosen such that the result of performing the operation 572 between the identity function value 590 and the data element is the value of the data element. For example, if the operation 572 was a multiply operation, then the identity function value 590 would be 1. Similarly, if the operation 572 was an add operation, the identity function value 590 would be 0. In this manner, the value of a data element can be selectively passed through to the logical destination register 430 by causing the corresponding MUX 575-577 to output the identity function value 590.

In the embodiments described above, the circuitry was hardwired such that the partial-width operation was performed on the least significant data element portion. It is appreciated that the operation may be performed on a different data element portions than illustrated. Also, as described above, the data elements to be operated upon may be made to be software configurable by coupling all of the operations to a MUX or the like, rather than simply a subset of the operations as depicted in Figures 5A-5C. Further, while pass through and clearing logic are described as two options for treating resulting data elements corresponding to operations that are to be disregarded, alternative embodiments may employ other techniques. For example, a QNaN may be

input as one of the operands to an operation whose result is to be disregarded. In this manner, arithmetic operations compliant with the IEEE 754 standard, IEEE std. 754-1985, published March 21, 1985, will propagate a NaN through to the result without triggering an arithmetic exception.

While no apparent speed up would be achieved in the embodiments described above since the full-width of the operands can be processed in parallel, it should be appreciated that power consumption can be reduced by shutting down those of the operations whose results will be disregarded. Thus, significant power savings may be achieved. Additionally, with the use of QNaNs and/or identity function values a predictable exception model may be maintained by preventing exceptions from being triggered by data elements that are not part of the partial-width packed data operation. Therefore, reported exceptions are limited to those raised in connection with the data element(s) upon which the partial-width packed data operation purports to operate.

Figure 6 illustrates a current processor implementation of an arithmetic logic unit (ALU) that can be used to execute full-width packed data instructions. The ALU of Figure 6 includes the circuitry necessary to perform operations on the full width of the operands (i.e., all of the data elements). Figure 6 also shows that the ALU may contain one or more different types of execution units. In this example, the ALU includes two different types of execution units for respectively performing different types of operations (e.g., certain ALUs use separate units for performing ADD and MUL operations). The ADD execution unit and the MUL execution unit are respectively capable of operating as four separate ADD execution units and four separate MUL execution units. Alternatively, the ALU may contain one or more Multiply Accumulate (MAC) units, each capable of performing more than a single type of operation. While the following examples assume the use of ADD and MUL execution units and floating point operations, it is appreciated other execution units such as MAC and/or integer operations may also be used. Further, it

may be preferable to employ a partial-width implementation (e.g., an implementation with less than a one-to-one correspondence between execution units and data elements) and additional logic to coordinate reuse of the execution units as described below.

Partial-Width Hardware and "Staggered Execution"

Figures 7A-7B conceptually illustrate a full-width packed data operation and a partial-width packed data operation being performed in a "staggered" manner, respectively. "Staggered execution" in the context of this embodiment refers to the process of dividing each of an instruction's operands into separate segments and sequentially processing each segment using the same hardware. The segments are sequentially processed by introducing a delay into the processing of the subsequent segments. As illustrated in Figures 7A-7B, in both cases, the packed data operands are divided into a "high order segment" (data elements 3 and 2) and a "low order segment" (data elements 1 and 0). In the example of Figure 7A, the low order segment is processed while the high order segment is delayed. Subsequently, the high order segment is processed and the full-width result is obtained. In the example of Figure 7B, the low order segment is processed, while whether the high order data segment is processed depends on the implementation. For example, the high order data segment may not need to be processed if the corresponding result is to be zeroed. Additionally, it is appreciated that if the high order data segment is not processed, then both the high and low order data segments may be operated upon at the same time. Similarly, in a full-width implementation (e.g., an implementation with a one-to-one correspondence between execution units and data elements) the high and low order data segments may be processed concurrently or as shown in Figure 7A.

Additionally, although the following embodiments are described as having only ADD and MUL execution units, other types of execution units such as MAC units may also be used.

While there are a number of different ways in which the staggered execution of instructions can be achieved, the following sections describe two exemplary embodiments to illustrate this aspect of the invention. In particular, both of the described exemplary embodiments receive the same macro instructions specifying logical registers containing 128 bit operands.

In the first exemplary embodiment, each macro instruction specifying logical registers containing 128 bit operands causes the full-width of the operands to be accessed from the physical registers. Subsequent to accessing the full-width operands from the registers, the operands are divided into the low and high order segments (e.g., using latches and multiplexers) and sequentially executed using the same hardware. The resulting half-width results are collected and simultaneously written to a single logical register.

In contrast, in the second exemplary embodiment each macro instruction specifying logical registers containing 128 bit operands is divided into at least two micro instructions that each operate on only half of the operands. Thus, the operands are divided into a high and low order segment and each micro instruction separately causes only half of the operands to be accessed from the registers. This type of a division is possible in a SIMD architecture because each of the operands is independent from the other. While implementations of the second embodiment can execute the micro instructions in any order (either an in order or an out of order execution model), the micro instructions respectively cause the operation specified by the macro instruction to be independently or separately performed on the low and high order segments of the operands. In addition, each micro instruction causes half of the resulting operand to be written into the single destination logical register specified by the macro instruction.

While embodiments are described in which 128 bit operands are divided into two segments, alternative embodiments could use larger or smaller operands and/or divide those operands into more than two segments. In

addition, while two exemplary embodiments are described for performing staggered execution, alternative embodiments could use other techniques.

First Exemplary Embodiment Employing "Staggered Execution"

Figure 8A conceptually illustrates circuitry within a processor according to a first embodiment that accesses full width operands from the logical registers but that performs operations on half of the width of the operands at a time. This embodiment assumes that the processor execution engine is capable of processing one instruction per clock cycle. By way of example, assume the following sequence of instructions is executed: ADD X, Y; MUL A, B. At time T, 128-bits of X and 128-bits of Y are each retrieved from their respective physical registers via ports 1 and 2. The lower order data segments, namely the lower 64 bits, of both X and Y are passed into multiplexers 802 and 804 and then on to the execution units for processing. The higher order data segments, the higher 64 bits of X and Y are held in delay elements M1 and M2. At time T+1, the higher order data segments of X and Y are read from delay elements M1 and M2 and passed into multiplexers 802 and 804 and then on to the execution units for processing. In general, the delay mechanism of storing the higher order data segments in delay elements M1 and M2 allows N-bit ($N=64$ in this example) hardware to process $2N$ -bits of data. The low order results from the execution unit are then held in delay element M3 until the high order results are ready. The results of both processing steps are then written back to register file 800 via port 3. Recall that in the case of a partial-width packed data operation one or more data elements of the low or high order results may be forced to a predetermined value (e.g., zero, the value of a corresponding data element in one of X or Y, etc.) rather than the output of the ADD or MUL operation.

Continuing with the present example, at time T+1, the MUL instruction may also have been started. Thus, at time T+1, 128-bits of A and B may each have been retrieved from their respective registers via ports 1 and 2. The lower order data segments, namely the lower 64-bits, of both A and B may be passed

into multiplexers 806 and 808. After the higher order bits of X and Y are removed from delay elements M1 and M2 and passed into multiplexers 806 and 808, the higher order bits of A and B may be held in storage in delay elements M1 and M2. The results of both processing steps is written back to register file 800 via port 3.

Thus, according to an embodiment of the invention, execution units are provided that contain only half the hardware (e.g. two single precision ADD execution units and two single precision MUL execution units), instead of the execution units required to process the full width of the operands in parallel as found in a current processor. This embodiment takes advantage of statistical analysis showing that multimedia applications utilize approximately fifty percent ADD instructions and fifty percent MUL instructions. Based on these statistics, this embodiment assumes that multimedia instructions generally follow the following pattern: ADD, MUL, ADD, MUL, etc.. By utilizing the ADD and MUL execution units in the manner described above, the present embodiment provides for an optimized use of the execution units, thus enabling comparable performance to the current processor, but at a lower cost.

Figure 8B is a timing chart that further illustrates the circuitry of Figure 8A. More specifically, as illustrated in Figure 8B, when instruction "ADD X, Y" is issued at time T, the two ADD execution units first perform ADDs on the lower order data segments or the lower two packed data elements of Figure 1, namely X_0Y_0 and X_1Y_1 . At time T + 1, the ADD operation is performed on the remaining two data elements from the operands, by the same execution units, and the subsequent two data elements of the higher order data segment are added, namely X_2Y_2 and X_3Y_3 . While the above embodiment is described with reference to ADD and MUL operations using two execution units, alternate embodiments may use any number of execution units and/or execute any number of different operations in a staggered manner.

According to this embodiment, 64-bit hardware may be used to process 128-bit data. A 128-bit register may be broken into four 32-bit elements, each of

which represents a separate 32-bit value. At time T, the two ADD execution units perform ADDs first on the two lower 32-bit values, followed by an ADD on the higher 32-bit values at time T+1. In the case of a MUL operation, the MUL execution units behave in the same manner. This ability to use currently available 64-bit hardware to process 128-bit data represents a significant cost advantage to hardware manufacturers.

As described above, the ADD and MUL execution units according to the present embodiment are reused to reexecute a second ADD or MUL operation at a subsequent clock cycle. Of course, in the case of a partial-width packed data instruction, the execution units are reused but the operation is not necessarily reexecuted since power to the execution unit may be selectively shut down. At any rate, as described earlier, in order for this re-using or "staggered execution" to perform efficiently, this embodiment takes advantage of the statistical behavior of multimedia applications.

If a second ADD instruction follows a first ADD instruction, the second ADD may be delayed by a scheduling unit to allow the ADD execution units to complete the first ADD instruction, or more specifically on the higher order data segment of the first ADD instruction. The second ADD instruction may then begin executing. Alternatively, in an out-of-order processor, the scheduling unit may determine that a MUL instruction further down the instruction stream may be performed out-of-order. If so, the scheduling unit may inform the MUL execution units to begin processing the MUL instruction. If no MUL instructions are available for processing at time T+1, the scheduler will not issue an instruction following the first ADD instruction, thus allowing the ADD execution units time to complete the first ADD instruction before beginning the second ADD instruction.

Yet another embodiment of the invention allows for back-to-back ADD or MUL instructions to be issued by executing the instructions on the same execution units on half clock cycles instead of full clock cycles. Executing an instruction on the half clock cycle effectively "double pumps" the hardware,

Although for simplicity only a single macro instruction is shown as being decoded at a time in the embodiment depicted, in alternative embodiments multiple macro instructions may be decoded concurrently. Also, it is appreciated that micro instruction replication has broader applicability than that illustrated by the above embodiment. For example, in a manner similar to that described above, full-width and partial-width packed data macro instructions may be decoded by the same decoder. If a prefix is used to distinguish full-width and partial width packed data macro instructions, the decoder may simply ignore the prefix and decode both types of instructions in the same manner. Then, the appropriate bits in the resulting micro operations may be modified to selectively enable processing for either all or a subset of the data elements. In this manner, full-width packed data micro operations may be generated from partial-width packed data micro operations or vice versa, thereby reducing complexity of the decoder.

Thus, a method and apparatus for efficiently executing partial-width packed data instructions are disclosed. These specific arrangements and methods described herein are merely illustrative of the principles of the invention. Numerous modifications in form and detail may be made by those of ordinary skill in the art without departing from the scope of the invention. Although this invention has been shown in relation to a particular preferred embodiment, it should not be considered so limited. Rather, the invention is limited only by the scope of the appended claims.

CLAIMS

What is claimed is:

1. A processor comprising:
a plurality of registers;
a register renaming unit coupled to the plurality of registers to provide
an architectural register file to store packed data operands, each
of said packed data operands having a plurality of data elements;
and
a decoder, coupled to said register renaming unit, to decode a first and
second set of instructions that each specify one or more registers
in the architectural register file, each instruction in the first set of
instructions specifying operations on all of the data elements
stored in the specified one or more registers, each of the second
set of instructions specifying an operation on only a subset of
data element stored in the specified one or more registers.
2. The processor of claim 1, wherein the subset of data elements stored in a
specified one or more registers comprises corresponding least significant
data elements.
3. The processor of claim 1, further comprising an execution unit to
selectively perform a specified operation on one or more data elements
in the specified one or more registers depending upon which of the first
or second set of instructions the specified operation is associated.
4. The processor of claim 3, wherein the execution unit further comprises a
plurality of multiplexers to select between a result of the specified
operation and a predetermined value.
5. The processor of claim 3, wherein the execution unit further comprises a
plurality of multiplexers to select between a data element of the one or

more data elements and an identity function for input to the specified operation.

6. A method of performing packed data operations comprising the steps of:
a decode unit receiving a single macro instruction specifying at least two logical registers from a packed data register file, the two logical registers respectively storing a first packed data operand and a second packed data operand having corresponding data elements;
the decode unit communicating one or more micro instructions derived from the single macro instruction to an execution unit which is configured to accommodate both full-width packed data operations and partial-width packed data operations; and
responsive to the one or more micro instructions, the execution unit selectively operating on one or more pairs of corresponding data elements to obtain a packed data result by
performing an operation specified by the single macro instruction on a first subset of corresponding data elements to obtain a first portion of the packed data result, and
setting a second portion of the packed data result to one or more predetermined values.
7. The method of claim 6, wherein the step of setting a second portion of the packed data result to one or more predetermined values further comprises the step of passing through values of data elements from the first packed data operand to the second portion of the packed data result.
8. The method of claim 6, wherein the step of setting a second portion of the packed data result to one or more predetermined values further

comprises the step of clearing data elements of the second portion of the packed data result.

9. The method of claim 6, wherein the step of setting a second portion of the packed data result to one or more predetermined values further comprises the step of setting data elements of the second portion of the packed data result to a not-a-number (NaN) indication.

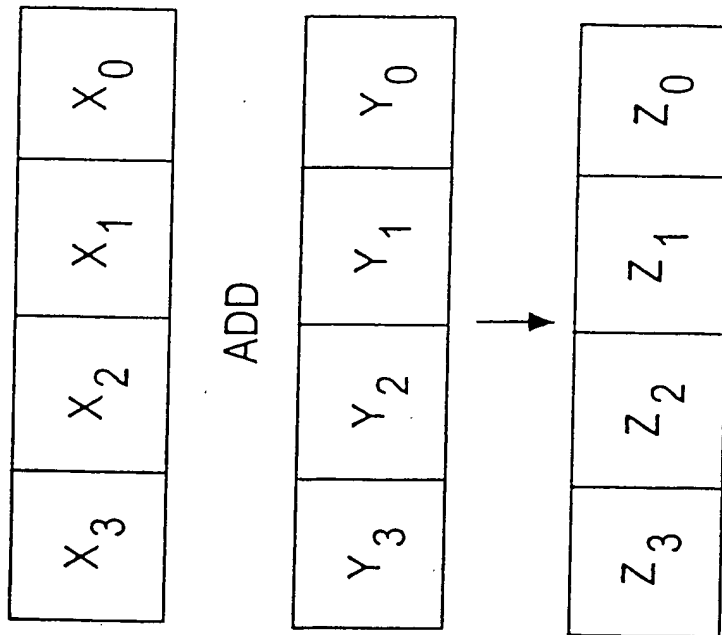


FIG. 1
(PRIOR ART)

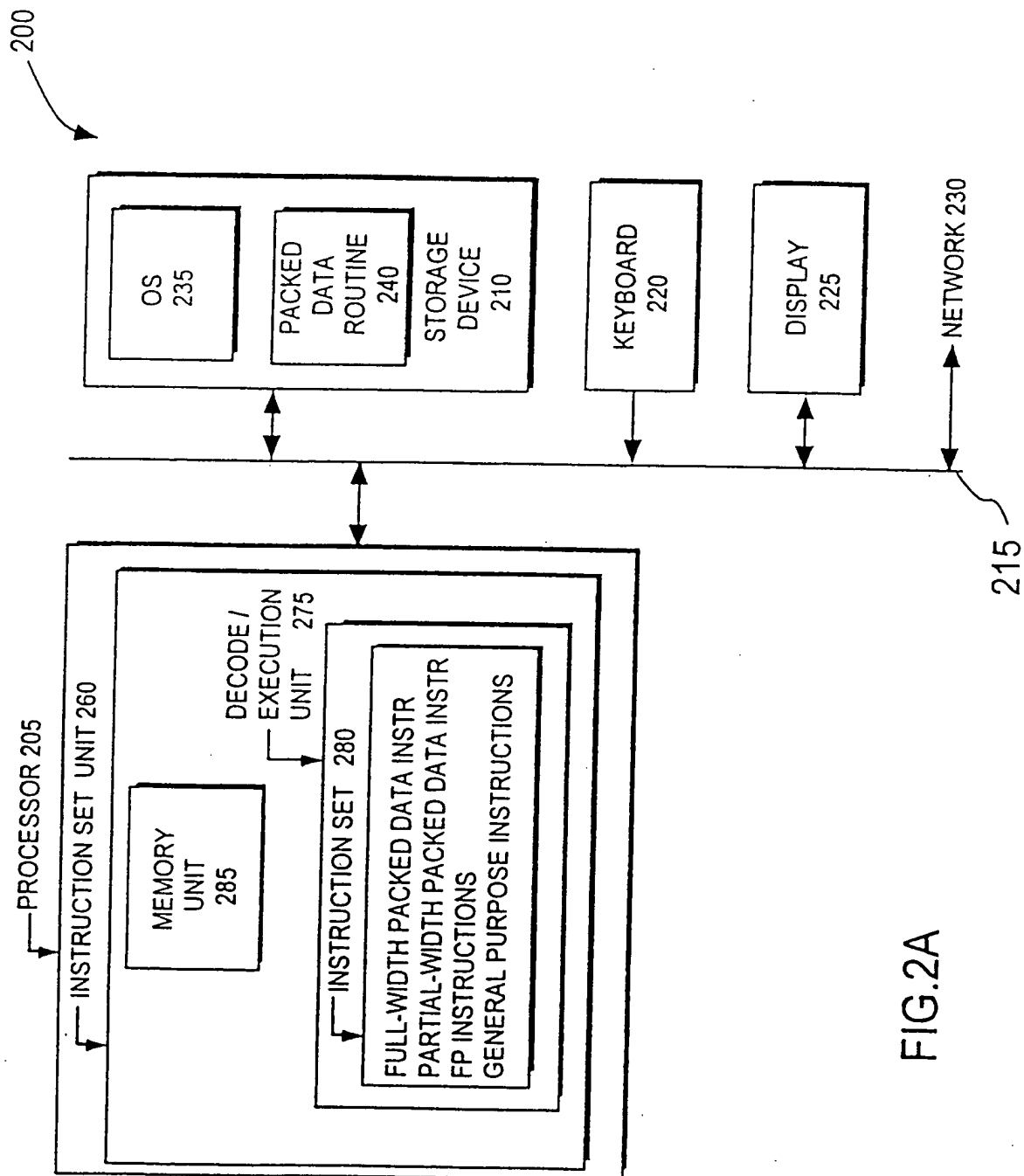


FIG.2A

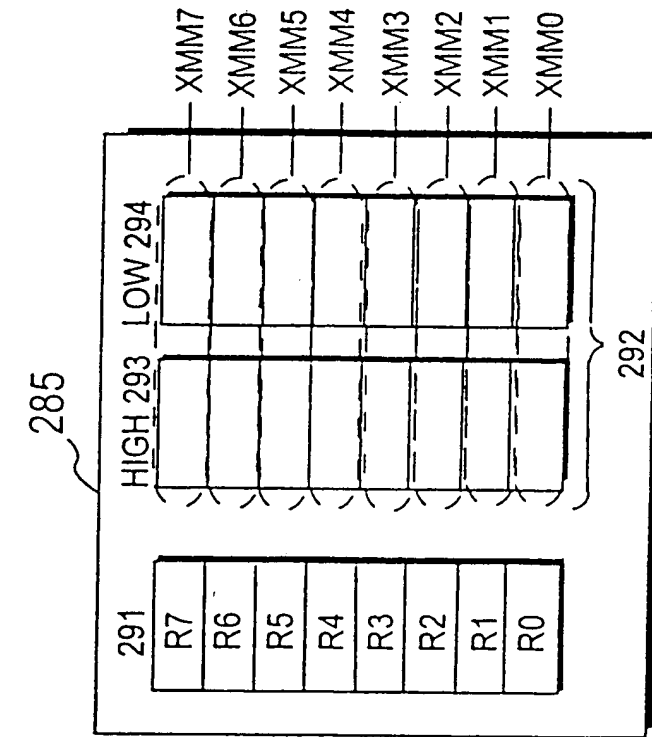


FIG. 2C

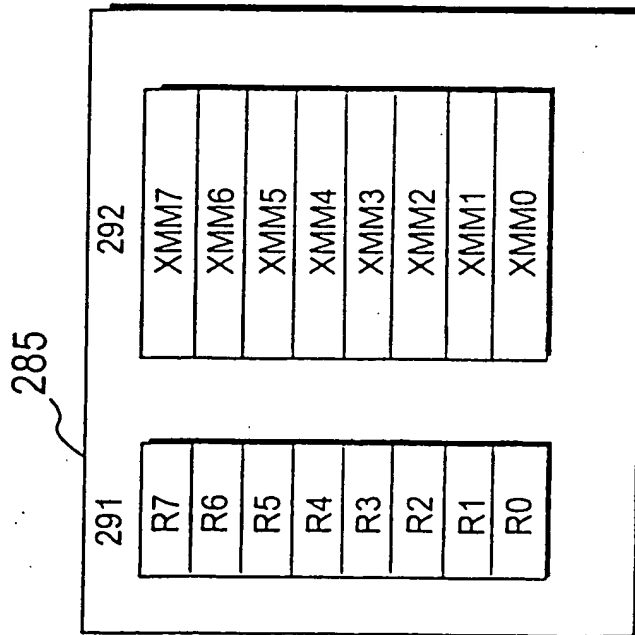


FIG. 2B

4/13

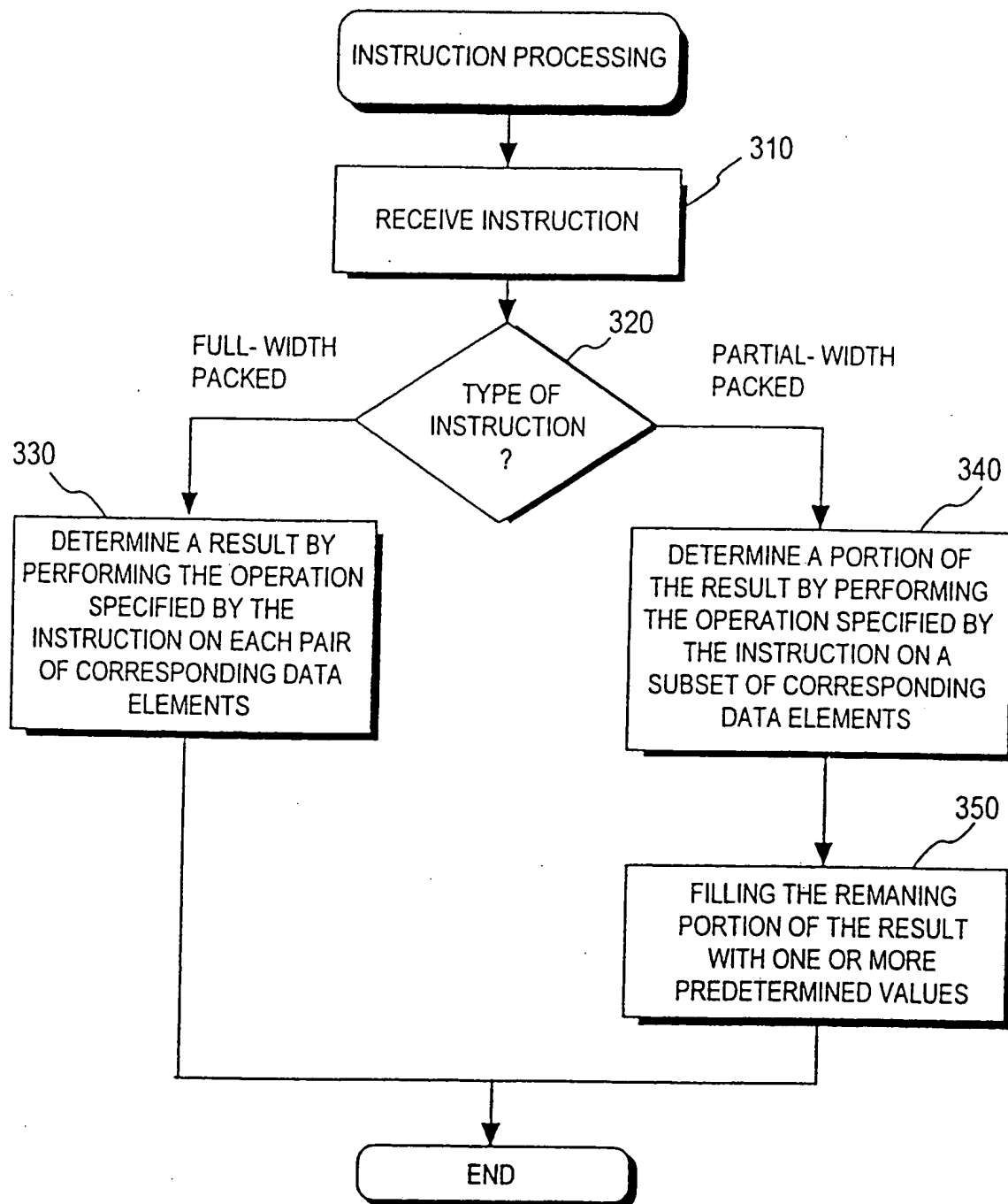


FIG. 3

5/13

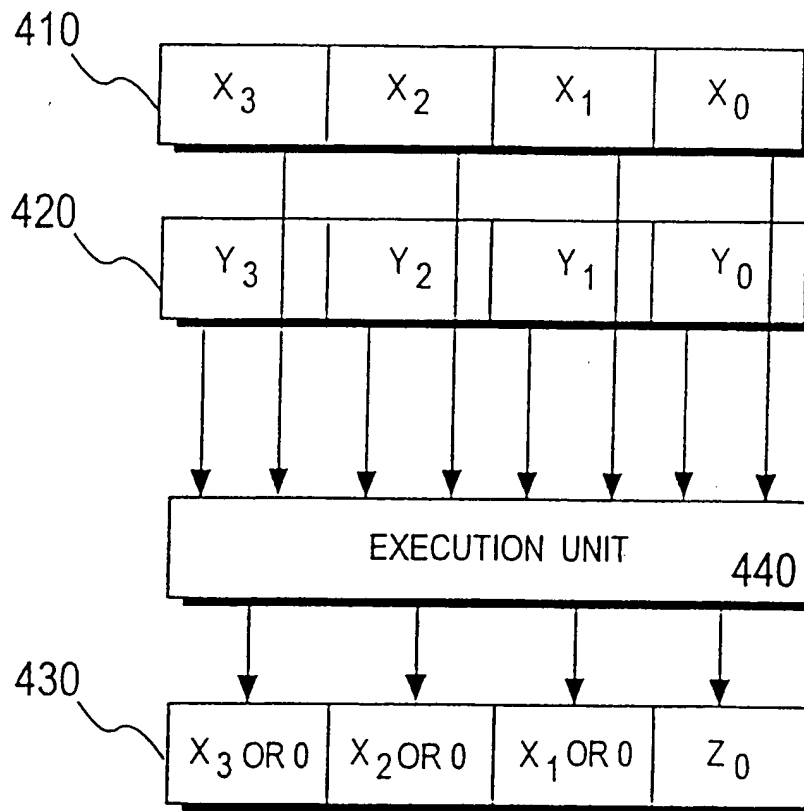
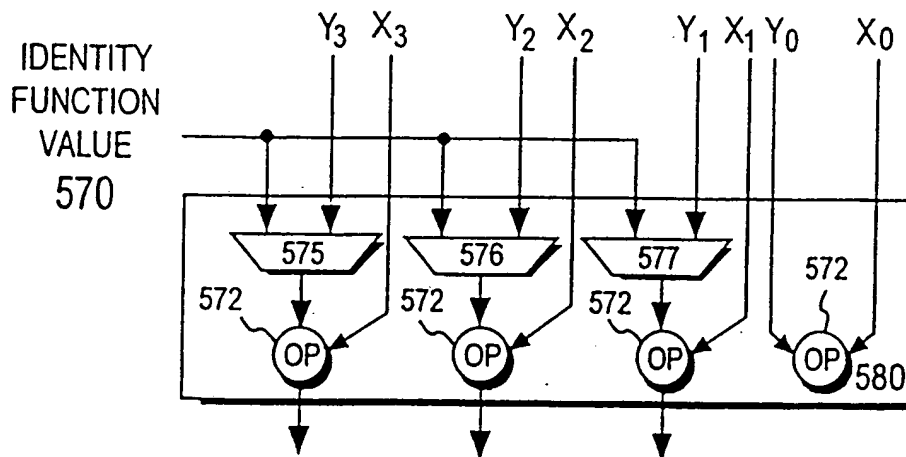
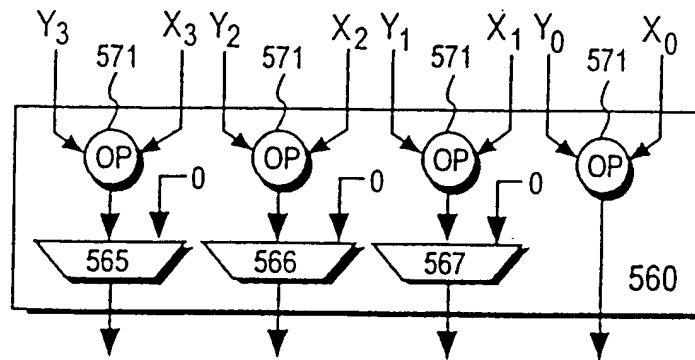
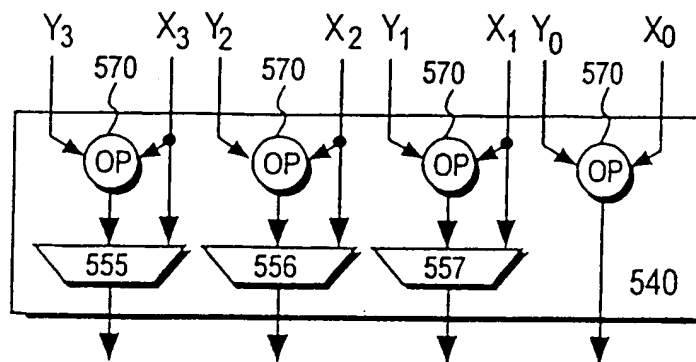


FIG. 4

6/13



7/13

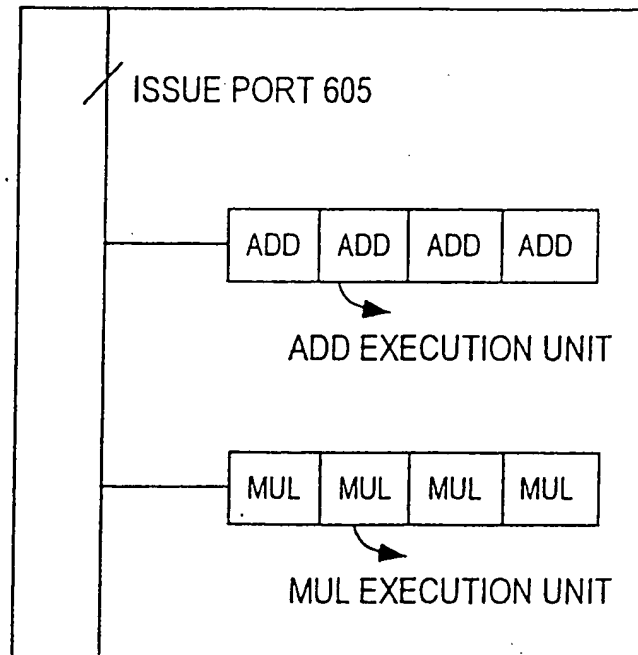


FIG. 6

8/13

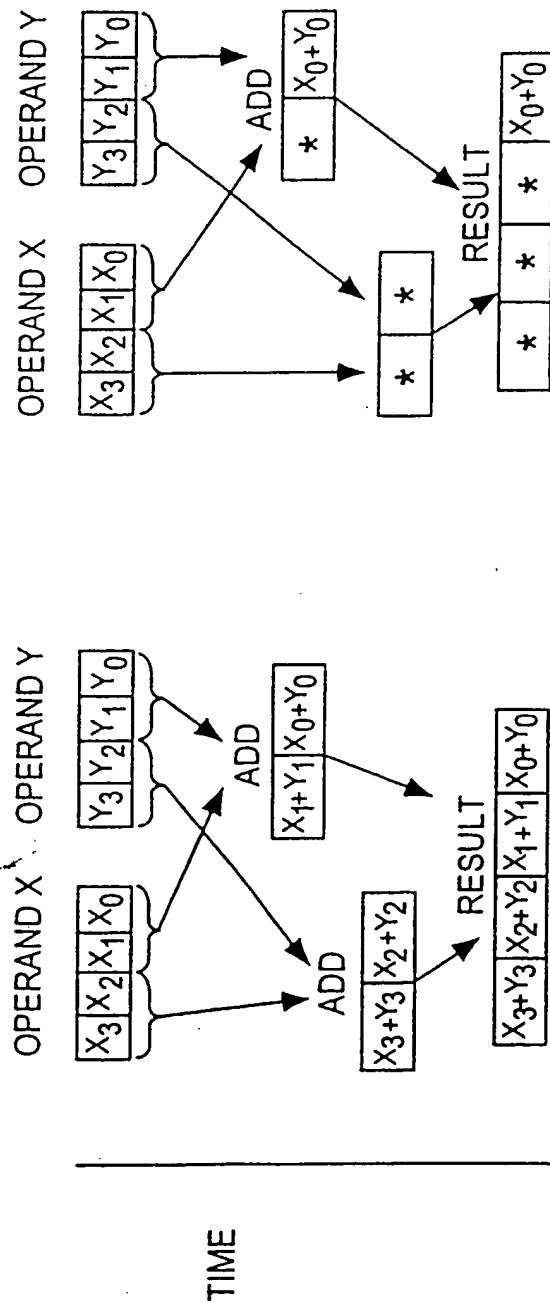


FIG. 7A

* = CORRESPONDING DATA ELEMENT IN X OR Y,
NaN, \emptyset , OR OTHER PREDETERMINED VALUE

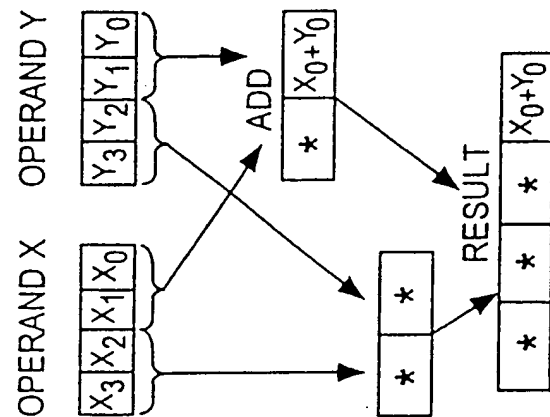


FIG. 7B

9/13

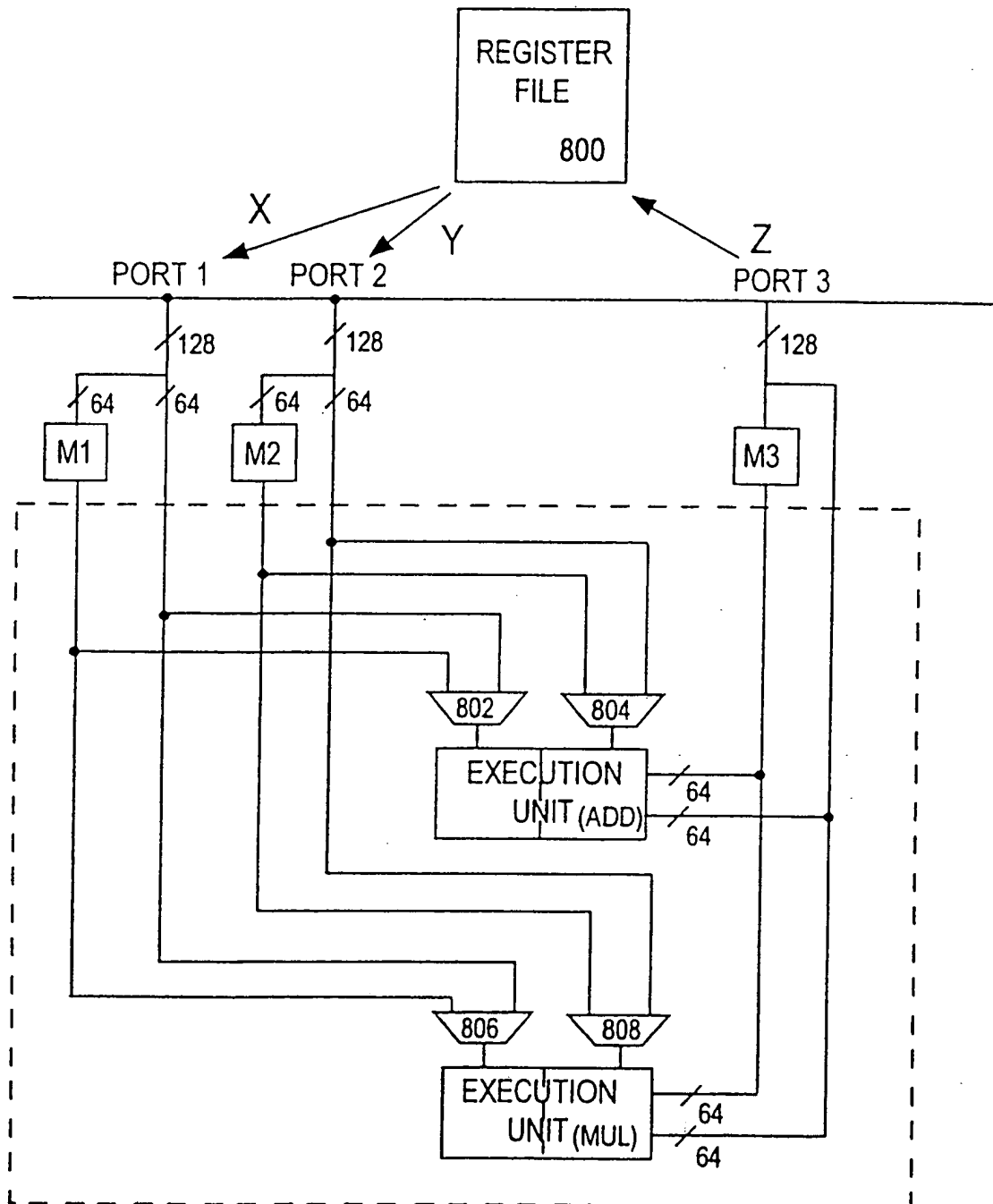


FIG. 8A

10/13

TIME	128-BIT INSTRUCTION	PERFORMED ON 64-BIT DATA
T	ADD X, Y	ADD X ₀ Y ₀ ADD X ₁ Y ₁
T + 1		ADD X ₂ Y ₂ ADD X ₃ Y ₃
T + 1	MUL X, Y	MUL X ₀ Y ₀ MUL X ₁ Y ₁
T + 2		MUL X ₂ Y ₂ MUL X ₃ Y ₃

FIG.8B

11/13

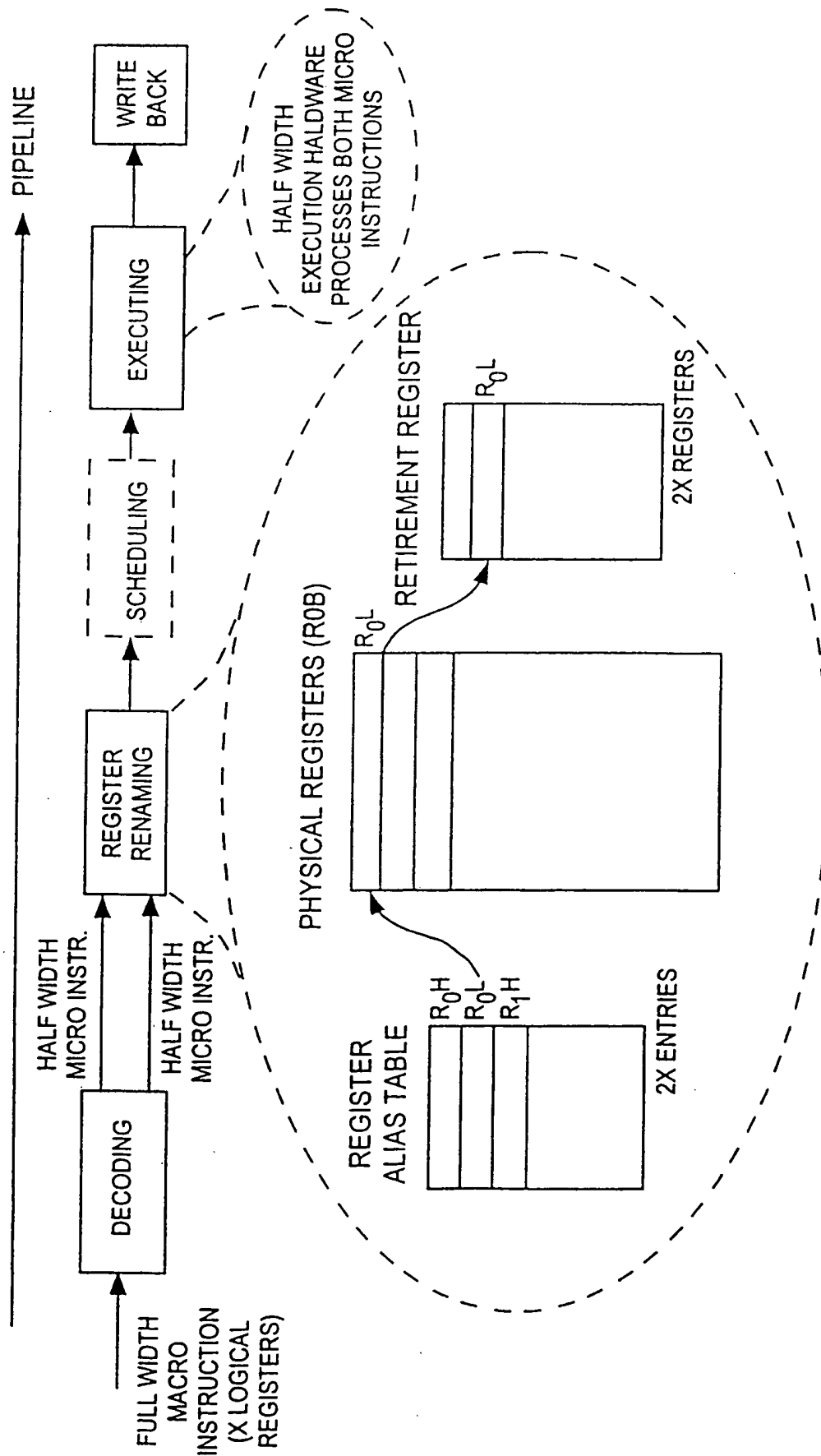


FIG.9

12/13

TIME	128-BIT INSTRUCTION	64-BIT INSTRUCTION
T	ADD X, Y	ADD X _L , Y _L
T+N		ADD X _H , Y _H

FIG. 10

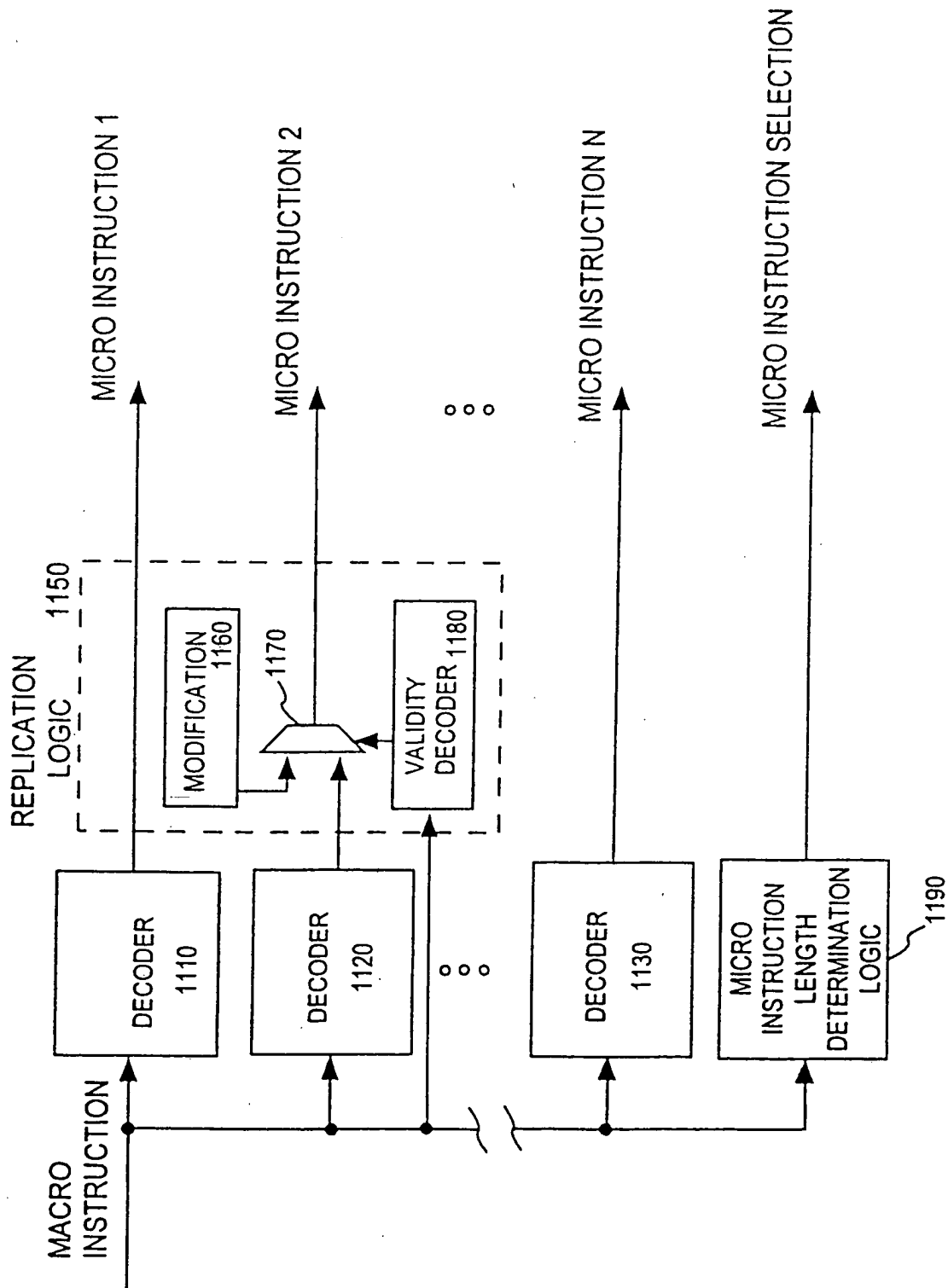


FIG. 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US99/04718

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : G06F 9/38, 9/34

US CL : 712/200, 712/23, 712/204

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 712/200, 712/23, 712/204

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS

search terms: rename, decode, packed, partial width, multiplexer

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X,P	US, 5,852,726 A (LIN et al) 22 December 1998 col. 36, line 62-col. 38, line 7, and fig. 10.	1-3,6-9
X	US, 5,673,427 A (BROWN et al) 30 September 1997, col. 10 lines 31-41 and col. 2, line 26-col. 3, line 33.	4,5
A	US, 4,890,218 A (BRAM) 26 December 1989, col. 1, lines 47-67.	1,6

☐ Further documents are listed in the continuation of Box C.
 ☐ See patent family annex.

•	Special categories of cited documents:	•T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
•A	document defining the general state of the art which is not considered to be of particular relevance		
•B	earlier document published on or after the international filing date	•X	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
•L	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	•Y	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
•O	document referring to an oral disclosure, use, exhibition or other means		
•P	document published prior to the international filing date but later than the priority date claimed	•A	document member of the same patent family

Date of the actual completion of the international search

10 JUNE 1999

Date of mailing of the international search report

28 JUN 1999

 Name and mailing address of the ISA/US
 Commissioner of Patents and Trademarks
 Box PCT
 Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

Eric Coleman

Telephone No. (703) 305-9674

Joni Hill

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☒ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.